Journal of Computer and Communications, 2026, 14(3), 71-90
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Seeing the Walk: Vision Transformers for Accurate Human Gait Recognition

Nouf Nayish M. Alghamdi[1], Osamah A. M. Ghaleb[2]

[1]Technical and Vocational Training Corporation (TVTC), Digital Technical College for Girls in Tabuk, Tabuk, Saudi Arabia
[2]Department of Computer Science, Fahad bin Sultan University, Tabuk, Saudi Arabia
Email: noufa3@tvtc.gov.sa, oghaleb@fbsu.edu.sa

How to cite this paper: Alghamdi, N.N.M. and Ghaleb, O.A.M. (2026) Seeing the Walk: Vision Transformers for Accurate Human Gait Recognition. *Journal of Computer and Communications*, **14**, 71-90.
https://doi.org/10.4236/jcc.2026.143005

Received: January 29, 2026
Accepted: March 13, 2026
Published: March 16, 2026

Abstract

Human gait recognition (HGR) is a non-invasive biometric modality that is applicable in mass-scale surveillance and security systems. Nevertheless, HGR systems are currently susceptible to covariate variables, including viewpoint, clothing, and the carrying conditions. Recent deep learning methods, especially convolutional neural networks, have also enhanced recognition performance, though at the cost of modeling global spatiotemporal interactions and with high training data requirements. This study presents a transfer learning-based HGR model that relies on Vision Transformer (ViT) models to harness self-attention mechanisms to achieve strong representation of global features. ViT-B/16 and ViT-L/32 are two pre-trained transformer models that were trained on gait image sequences on the CASIA-B dataset. The framework was evaluated across four viewing angles (0˚, 18˚, 36˚, and 54˚) under varying covariate conditions. Training and testing accuracy and loss metrics were used as performance metrics at the learning rate of 0.001 and 0.0001. Furthermore, experiments were conducted for both frontal-view and cross-view analysis. The results indicate that transformer-based models are capable of achieving strong recognition performance. ViT-L/32 achieved the highest average testing accuracy at 87.87 percent, followed closely by ViT-B/16 with 86.99 percent. Both models outperformed several recently proposed HGR approaches. The attention-based architecture successfully extracts discriminative gait images across image patches and is more robust to viewpoint and appearance changes as well as it consumes less computational costs due to pre-trained models. These results highlight the usefulness of Vision Transformers as an effective and precise alternative to traditional deep learning methods of recognizing human gait in biometric applications.

Keywords

Human Gait Recognition, Vision Transformer Models, Gait Pattern

1. Introduction

By facilitating reliable identification and authentication of people using physiological and behavioural characteristics, biometric systems have become essential features of contemporary security, surveillance, and access control systems. Traditional biometric modalities, such as fingerprints, faces, iris, and retinas demonstrate high recognition when used under controlled settings, but they are susceptible to occlusions, face impersonation, privacy issues, and require the cooperation of users. Conversely, human gait recognition (HGR) has become a viable behavioural biometric method as a non-invasive biometric identification tool and because data are collected remotely with ease in a regular camera system [1]. The innate distinctiveness of the walking style of an individual can be recognized even under low-resolution data, and in uncontrolled settings. Therefore, gait is especially appropriate in massive surveillance and forensic applications [2].

Irrespective of these strengths, the efficiency of HGR systems is susceptible to a number of covariate variables, such as changes in viewing angle, clothing, walking speed, illumination conditions, and carrying object [3]. These covariates add intra-class variability and inter-class similarity, which severely compromises recognition accuracy. To overcome these challenges, several techniques have been proposed in the last ten years which can be categorized into model-based and model-free techniques. Model-based approaches rely on direct models of human anatomy and joint dynamics with the capability to investigate movements in detail, although they are often computationally intensive and sensitive to noise. Alternatively, model-free methods frequently utilize silhouette-based characteristics such as Gait Energy Images (GEIS) to represent spatiotemporal motion patterns that are computationally inexpensive but sensitive to both appearance variations and viewpoint variation [4].

The recent advances in machine learning and deep learning have improved the gait recognition performance to a significant level. Convolutional Neural Networks (CNNs), recurrent models and autoenergizer based models have been widely applied in the learning of discriminative gait features on image sequences or silhouette templates [5]. These techniques have proven quite effective in addressing the moderate covariate variations but CNN-based models typically use local receptive fields and pooling mechanisms, which may limit their ability to capture long-range spatial responses to model complex gait dynamics [6]. In addition, deep CNN models typically require bulk data and training, as well as higher computational costs [7].

Recent efforts in attention-based models, especially Vision Transformers (ViTs) have shown impressive results on a wide range of computer vision problems, including image classification, object detection and action recognition [8]. ViTs, unlike CNNs, use large patches to divide pictures and use self-attention models that capture the global contextual associations between the whole picture [9]. This ability allows ViTs to learn the long-range and subtle discriminative patterns that are important in biometric recognition tasks. The self-attention mechanism in the

context of gait analysis provides a possibility to concentrate on the salient areas of motion without being influenced by appearance-related differences [10].

Contemporary studies have started examining transformer-based architecture in biometric and gait-related uses. Hybrid CNN-transformer networks have been introduced to complement local feature extraction with global attention modelling, with enhanced robustness to viewpoint and clothing variation [11] [12]. Pure transformer-based architectures have also exhibited state-of-the-art results in learning spatiotemporal motion representations, especially when combined with transfer learning methods exploiting large-scale pre-trained models [12]. These advances imply that Vision Transformers could offer an effective alternative to traditional deep learning methods in HGR, particularly in settings with strong covariate variance [13] [14].

Although there is remarkable progress in HGR with both classical machine learning and deep learning methods, several limitations remain [15]. The model-based approaches tend to be computationally complex and noise sensitive whereas model-free silhouette-based approaches remain susceptible to changes in clothing, viewpoint, illumination and carrying conditions [16]. Convolutional neural networks and recurrent architectures have enhanced the learner capacity of features, but they cannot model the global spatiotemporal features due to their dependence on the local receptive fields and generally need a large labelled dataset, which increases computational demands [17]. Vision Transformer models provide a potential alternative as they use self-attention to learn global contextual dependencies [14]. Nevertheless, their direct use in gait biometrics is restricted. The current study bridges this gap by offering a transfer learning based HGR framework using pre-trained ViT-B/16 and ViT-L/32 models to effectively learn discriminative gait representations on image sequences. The suggested methodology involves a multi-view assessment of the CASIA-B dataset, comparison with the state-of-the-art HGR-based methods, and optimisation over the variant of transformers, showing higher level of robustness and performance. This study combines self-attention worldwide with transfer learning to promote gait biometric recognition and make Vision Transformer a viable alternative to traditional CNN-based models.

2. Methods

2.1. Overview of the Proposed Framework

The proposed study presents an HGR model on Vision Transformer (ViT) architecture with a transfer learning modality. The general workflow involves gait frame detection on video sequences, image processing, dataset division, and transformer-based feature detection and classification. The complete architecture of the proposed system is illustrated in **Figure 1**, which outlines the sequential processing pipeline from data acquisition to final recognition.

2.2. Dataset and Data Acquisition

CASIA-B gait dataset, which is publicly available, was used to test the proposed

framework. The data set consists of gait sequences recorded on 124 participants in controlled conditions in an indoor environment at 11 viewing angles (0˚ to 180˚ with 18˚ intervals). All subjects completed ten walking sequences of which six of them were normal walking sequences (NM), two walking sequences with a carried bag (BG), and two walking sequences with a coat (CL), which were all common covariate conditions that influenced gait recognition. Video-to-image extraction process was used to convert video sequences into image frames. Six frames of every video sequence were uniformly sampled to ensure balance in representation and computational efficiency. The extracted images were all resized at 256 × 256 pixels in three colour channels. The resulting processed data was then further subdivided into training and testing groups so that the performance can be assessed impartially.

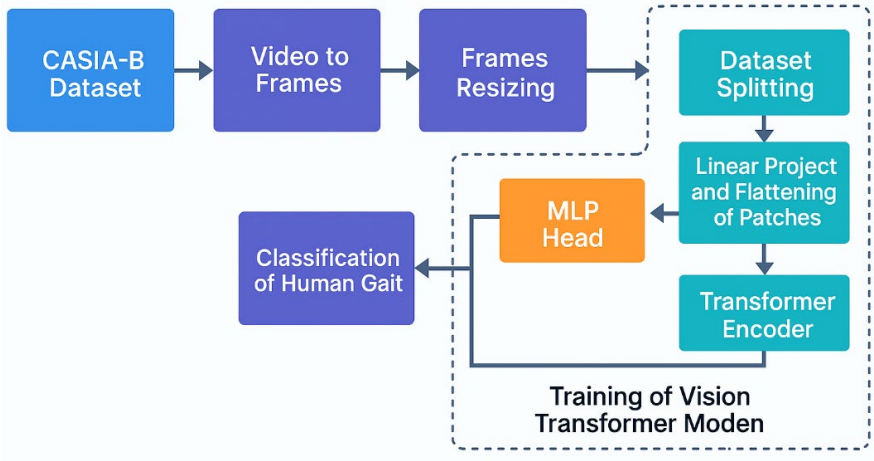

Figure 1. The architecture of the proposed HGR method.

2.3. Vision Transformer-Based Gait Recognition

The suggested framework employs Vision Transformer models to use self-attention mechanisms to learn the global representation of features. In contrast to traditional convolutional neural networks, where local spatial patterns are addressed, ViTs subdivide images into fixed patch sizes and operate on the sequences of the tokens, which allows to effectively model long-range effects. This study adopted two variants of transformers that were pre-trained.

ViT-B/16 and ViT-L/32, where the numbers indicate patch sizes used for image segmentation. Let the input image be represented as

$$I \in \mathbb{R}^{L\times C\times O}$$

where L, C, and O denote image height, width, and number of channels, respectively. The image is partitioned into non-overlapping patches, which are flattened to form patch vectors

$$X_p \in \mathbb{R}^{N\times(p\cdot O)}$$

where p represents the patch resolution and N denotes the total number of

patches, computed as:

$$N = \frac{L \times C}{p^2} \tag{1}$$

This patch-based representation significantly reduces computational complexity while preserving spatial information.

2.4. Linear Projection and Positional Encoding

Each flattened patch was linearly projected into an embedding space using a learnable projection matrix. To retain spatial relationships between patches, positional embeddings were added to the patch embeddings. A learnable classification token was appended to the sequence to enable global representation learning for classification.

The embedded token sequence was formulated as:

$$z = \left[x_{class}; x_1 E; x_2 E; \cdots; x_N E\right] \tag{2}$$

where E denotes the embedding matrix and x_{class} represents the class token. This combined sequence served as input to the transformer encoder.

2.5. Transformer Encoder Architecture

The transformer encoder was composed of several layers, each with a multi-head self-attention (MSA) module then a multi-layer perceptron (MLP) as in Figure 2. Residual connections and layer normalisation were added in order to enhance the stability of training and prevent the problem of vanishing gradients.

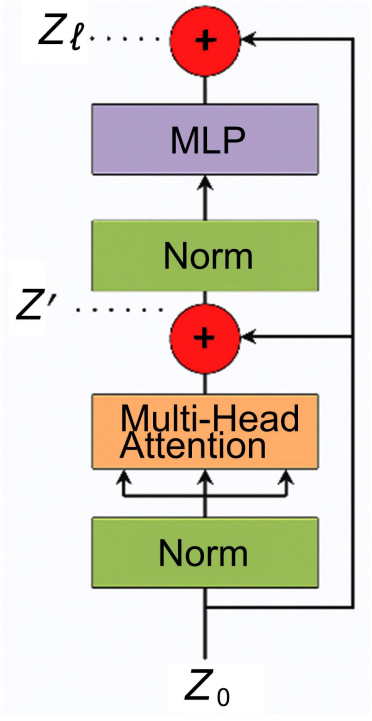

Figure 2. The architecture of the transformer encoder.

For the l^{th} encoder layer, the computations were expressed as:

$$z'_l = \mathrm{MSA}\left(\mathrm{LN}\left(z_{l-1}\right)\right) + z_{l-1}, l = 1, \cdots, L \tag{3}$$

$$z_l = \mathrm{MLP}\left(\mathrm{LN}\left(z'_l\right)\right) + z'_l \tag{4}$$

where LN denotes layer normalisation.

2.6. Multi-Head Self-Attention Mechanism

The MSA module performs parallel self-attention operations across multiple heads to capture diverse feature relationships. For each head, the input embeddings were projected to query (q), key (k), and value (v) matrices:

$$[q, k, v] = \left[zU_q, zU_k, zU_v\right] \tag{5}$$

where U_q, U_k, and U_v are learnable weight matrices.

The attention weights were computed using the scaled dot-product mechanism:

$$A = \mathrm{softmax}\left(\frac{qk^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{6}$$

The self-attention output was then obtained as:

$$SA(z) = Av \tag{7}$$

Outputs from all heads were concatenated and linearly transformed:

$$\mathrm{MSA}(z) = \left[SA_1(z); SA_2(z); \cdots; SA_k(z)\right] U_{msa} \tag{8}$$

as depicted in Figure 3.

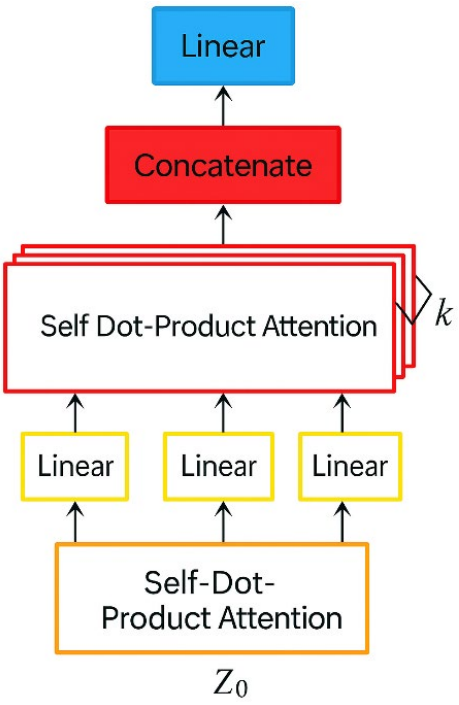

Figure 3. The architecture of MSA.

2.7. Multi-Layer Perceptron Block

The MLP block comprised two fully connected layers with Gaussian Error Linear Unit (GeLU) activation, as shown in **Figure 4**. GeLU assigns weights based on both magnitude and sign, enabling improved nonlinear representation learning compared to ReLU.

The MLP output was expressed as:

$$r_n = \mathrm{GeLU}\left(W_2 \cdot \mathrm{GeLU}\left(W_1 z_n + b_1\right) + b_2\right) \tag{9}$$

where z_n denotes the embedding of the n^{th} patch.

2.8. Classification and Model Training

The classifier head was replaced with a single linear layer with classes outputs followed by softmax with Cross-Entropy loss, ensuring compatibility with single-label multi-class recognition.

Model optimisation was performed using the Adam optimizer, and categorical cross-entropy loss was employed:

$$\mathcal{L} = -\sum_{t=1}^{T} y_t \log\left(\hat{y}_t\right) \tag{10}$$

where y_t represents the ground truth label, $\hat{y}_t$ denotes the predicted probability, and T is the number of samples.

Both ViT-B/16 and ViT-L/32 models were fine-tuned using learning rates of 0.001 and 0.0001, a batch size of 16, and 200 training epochs.

2.9. Experimental Setup and Evaluation

Experiments were conducted using four selected viewing angles of the CASIA-B dataset (0˚, 18˚, 36˚, and 54˚) to assess performance. Model performance was evaluated using training accuracy, testing accuracy, training loss, and testing loss.

The models were evaluated under three experimental setups: a 70:30 data split, a frontal view analysis, and a cross-view analysis. In the 70:30 configuration, the NM, BG, and CL classes of the CASIA-B dataset were included. Each class contained sequences from all subjects. This experiment was designed to examine how well the models handled variations in appearance.

A second set of experiments focused on subject recognition under different appearance conditions. Three evaluations were carried out: gallery versus NM, gallery versus CL, and gallery versus BG. The gallery set consisted of sequences NM-01 to NM-04, which were used for training. The corresponding test sets were NM-05 to NM-06, CL-01 to CL-02, and BG-01 to BG-02. These sequences allowed the system to be analyzed under normal walking, clothing variations, and carrying conditions.

The third experimental setup was intended to measure cross-view performance. The model was trained using sequences captured at the 000˚ view. It was then tested using sequences captured at the 180˚ view. This configuration introduced a significant change in viewpoint, providing insight into how effectively the model

generalized to unseen angles. All experiments were implemented in a Jupyter Notebook environment on a system equipped with an Intel Core i5 (9th generation) processor, 32 GB RAM, and an NVIDIA RTX 2060 GPU.

3. Results

3.1. Experimental Setup

The proposed Vision Transformer-based HGR framework was evaluated using four viewing angles of the CASIA-B dataset, namely 0˚, 18˚, 36˚, and 54˚. All experiments were implemented in a Jupyter Notebook environment using an Intel Core i5 (9th generation) processor, 32 GB RAM, 1.5 TB SSD storage, and an NVIDIA RTX 2060 GPU with 8 GB memory. Two pre-trained transformer models, ViT-B/16 and ViT-L/32, were fine-tuned using a batch size of 16 for 200 epochs. Learning rates of 0.001 and 0.0001 were examined to assess model sensitivity to optimisation settings.

Model performance was evaluated using training accuracy (Trn-ACR), testing accuracy (Tst-ACR), training loss (Trn-LS), and testing loss (Tst-LS).

3.2. Performance of ViT-B/16 at Learning Rate 0.001

The ViT-B/16 model was first evaluated using a learning rate of 0.001 across the selected CASIA-B viewing angles. At 0˚, the model achieved a training accuracy of 93.40% and a testing accuracy of 91.47%, with corresponding training and testing losses of 17.59% and 22.01%, respectively. At 18˚, training and testing accuracies decreased slightly to 89.91% and 86.80%, with losses of 26.82% and 34.27%. For the 36˚ angle, the model obtained a training accuracy of 89.22% and a testing accuracy of 86.28%, while training and testing losses were recorded as 28.15% and 34.56%. At 54˚, the performance further declined, with training and testing accuracies of 88.95% and 83.43%, and losses of 28.44% and 40.50%.

The detailed numerical results are presented in Table 1, while the corresponding training and validation curves are illustrated in Figure 4.

Table 1. Results for the ViT-B/16 model using CASIA-B at a learning rate of 0.001.

Angle (˚)	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
0	93.40	91.47	17.59	22.01
18	89.91	86.80	26.82	34.27
36	89.22	86.28	28.15	34.56
54	88.95	83.43	28.44	40.50

3.3. Performance of ViT-B/16 at Learning Rate 0.0001

When the learning rate was reduced to 0.0001, the ViT-B/16 model exhibited a noticeable decrease in recognition performance. At 0˚, the training accuracy was 89.98% and the testing accuracy was 87.86%, with losses of 29.34% and 32.10%, respectively.

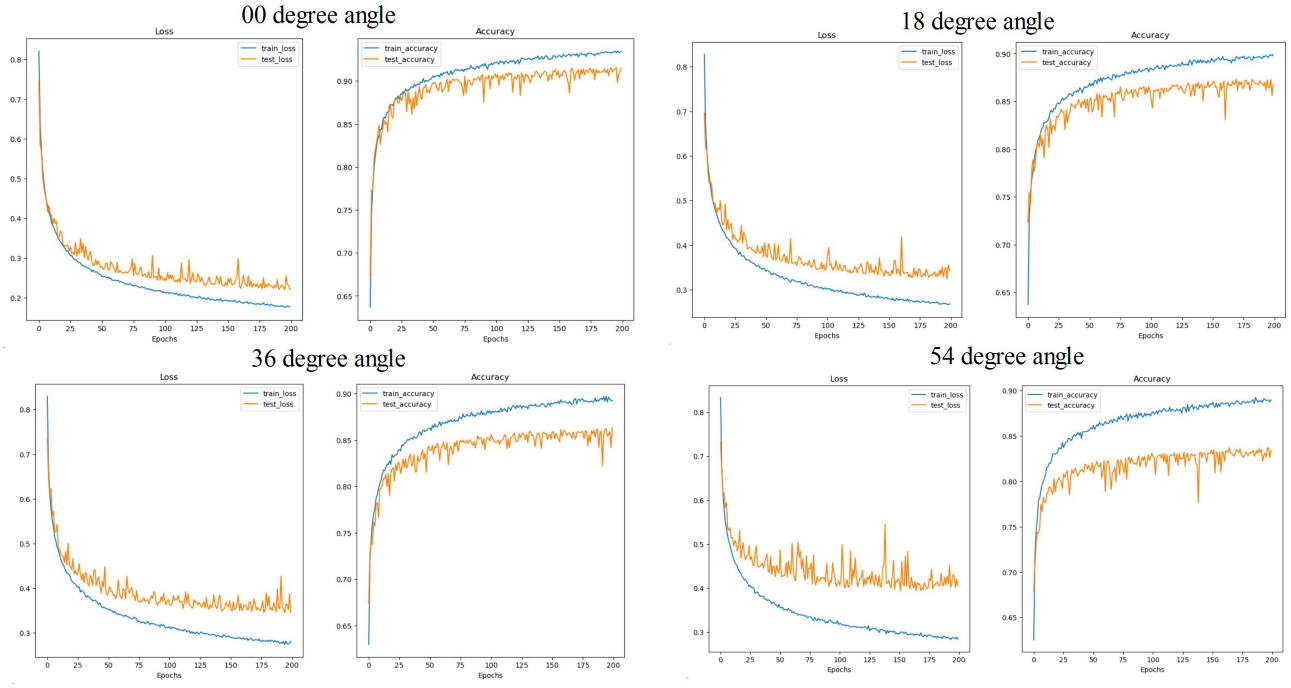

Figure 4. Training curves for the ViT-B/16 model using CASIA-B at a learning rate of 0.001.

At 18˚, performance dropped to 82.93% training accuracy and 81.15% testing accuracy. For 36˚, the model attained 81.20% training accuracy and 79.33% testing accuracy, with increased losses exceeding 50%. At 54˚, training and testing accuracies were recorded as 82.60% and 78.84%, respectively.

The complete results are summarised in **Table 2**, with training behaviour illustrated in **Figure 5**.

Table 2. Results for the ViT-B/16 model using CASIA-B at a learning rate of 0.0001.

Angle (˚)	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
0	89.98	87.86	29.34	32.10
18	82.93	81.15	42.99	47.05
36	81.20	79.33	51.62	50.90
54	82.60	78.84	41.18	50.90

3.4. Performance of ViT-B/16 for Frontal View Analysis Using 00˚ Angle of CASIA-B

For the frontal-view analysis using the 000˚ angle of the CASIA-B dataset, four sets were used in the experiments: the gallery view, NM, BG, and CL. The gallery view consisted of four sequences (NM-01 to NM-04), while NM-05 to NM-06, BG-01 to BG-02, and CL-01 to CL-02 were used as the evaluation sets. Each set contained gait sequences from all 124 subjects. The objective of this experiment was to investigate the model's ability to handle variations in appearance.

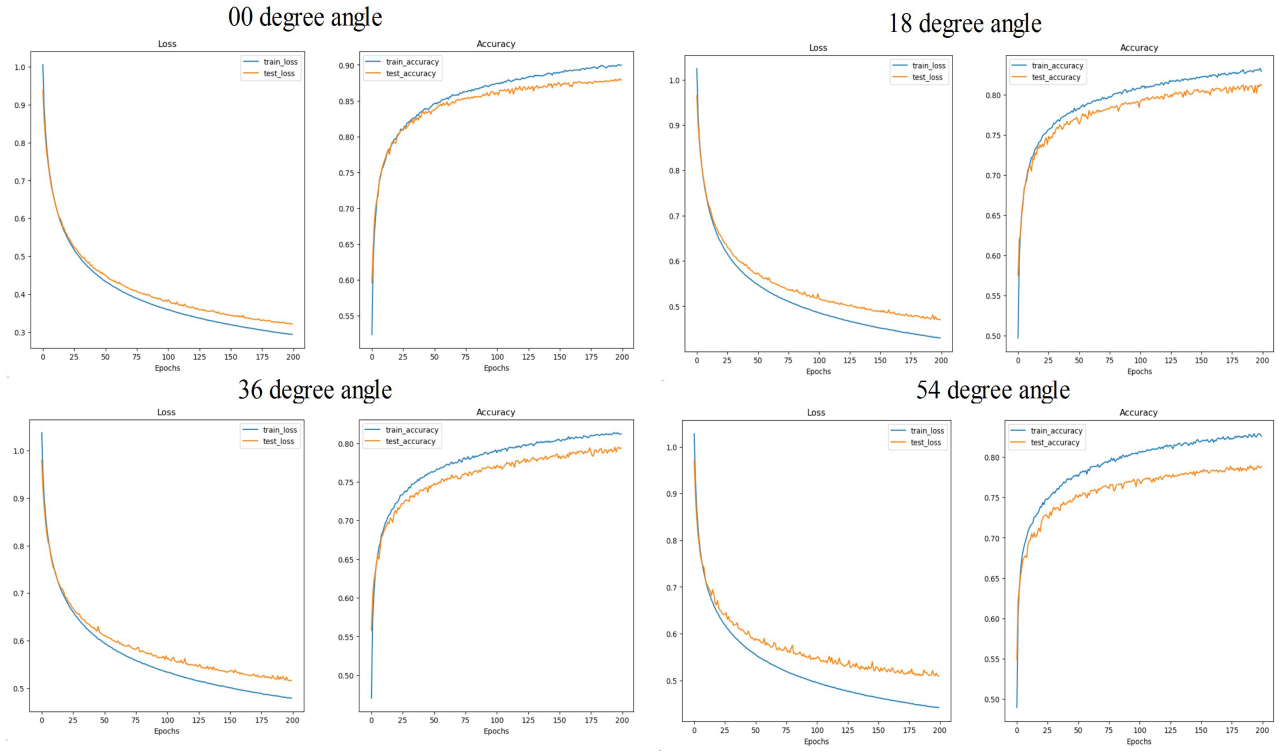

Figure 5. Training curves for the ViT-B/16 model using CASIA-B at a learning rate of 0.0001.

Using the gallery versus NM configuration, ViT-B/16 achieved a training accuracy of 100%, a test accuracy of 98.51%, a training loss of 0.00, and a test loss of 1.10. Under the gallery vs. CL setting, the model obtained training and test accuracies of 100% and 33.20%, respectively, with corresponding losses of 0.00 and 7.02. In the gallery versus BG evaluation, the training accuracy, test accuracy, training loss, and test loss were 100%, 98.43%, 0.00, and 2.10, respectively.

These results indicate that ViT-B/16 performs strongly on the NM and BG sets but exhibits a notable decline in performance on the CL set. The detailed results are provided in **Table 3**.

Table 3. Performance of ViT-B/16 for frontal view analysis using 00˚ angle of CASIA-B.

Variation	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
NM	100	98.51	0.00	1.1
CL	100	33.20	0.00	7.02
BG	100	98.43	0.00	2.10
Average	100	76.71	0.00	3.40

The training curves for all experiments using frontal view on ViT-B/16 are illustrated in **Figure 6**.

3.5. Performance of ViT-B/16 for Cross View Analysis

For the cross-view analysis of ViT-B/16, 70% of the gait sequences captured at the

000˚ view of the CASIA-B dataset were used for training, whereas 30% of the sequences captured at the 180˚ view were used for testing. The evaluation employed four metrics: training accuracy, test accuracy, training loss, and test loss. The model achieved a training accuracy of 100%, a test accuracy of 38.95%, a training loss of 0.00, and a test loss of 11.1123. The corresponding training curve for the cross-view evaluation using ViT-B/16 is presented in **Figure 7**.

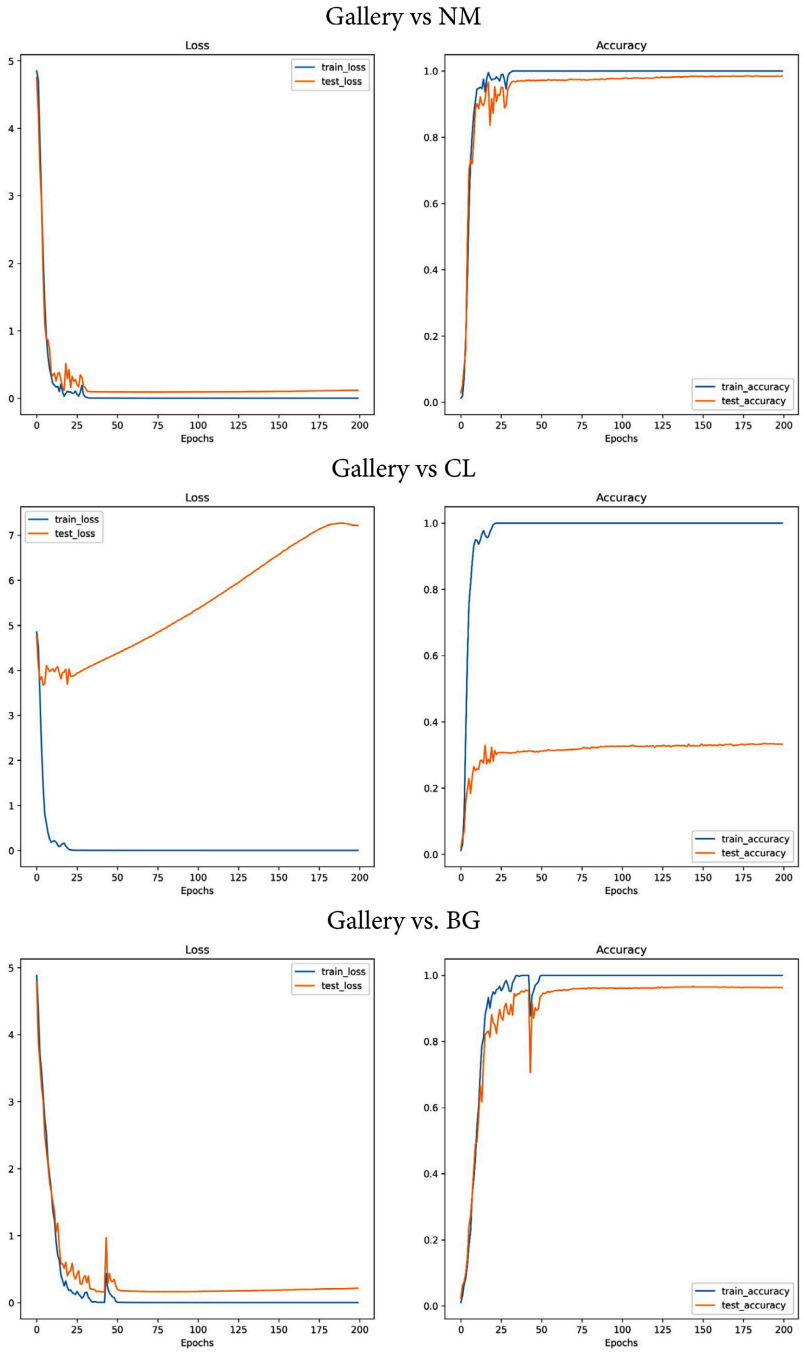

Figure 6. Training curves of ViT-B/16 for frontal view analysis using 00˚ angle of CASIA-B.

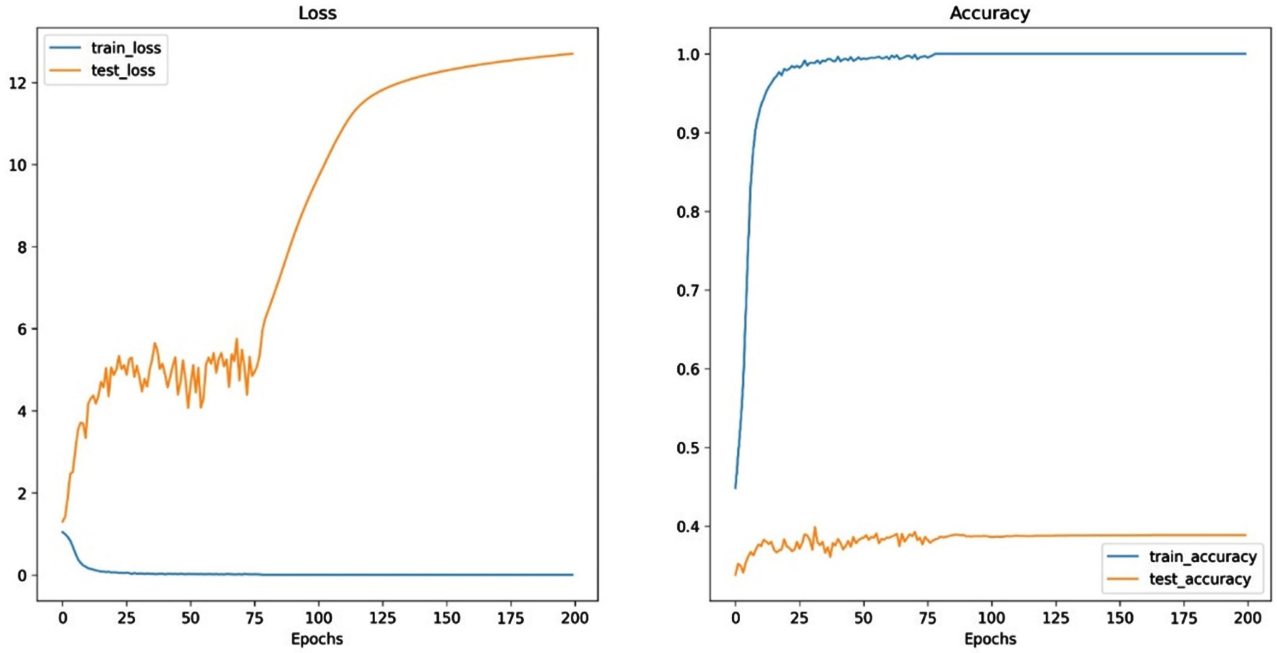

Figure 7. Training curves of ViT-B/16 for cross view analysis.

3.6. Performance of ViT-L/32 at Learning Rate 0.001

The ViT-L/32 model demonstrated improved recognition performance compared with ViT-B/16 when trained using a learning rate of 0.001. At 0˚, the model achieved a training accuracy of 96.33% and a testing accuracy of 94.02%, with reduced losses of 10.98% and 16.29%, respectively.

At 18˚, training and testing accuracies were 90.96% and 87.64%. At 36˚, the model obtained 89.94% training accuracy and 86.69% testing accuracy. At 54˚, performance decreased to 89.38% training accuracy and 83.12% testing accuracy.

The quantitative outcomes are shown in Table 4, and the training trends are illustrated in Figure 8.

Table 4. Results for the ViT-L/32 model using CASIA-B at a learning rate of 0.001.

Angle (˚)	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
0	96.33	94.02	10.98	16.29
18	90.96	87.64	23.83	32.64
36	89.94	86.69	26.44	32.04
54	89.38	83.12	26.74	40.05

3.7. Performance of ViT-L/32 at Learning Rate 0.0001

At the lower learning rate of 0.0001, ViT-L/32 achieved slightly reduced performance relative to its higher learning rate configuration. At 0˚, training and testing accuracies were 92.25% and 90.75%, respectively. Similar performance was observed at 18˚, with a testing accuracy of 90.69%.

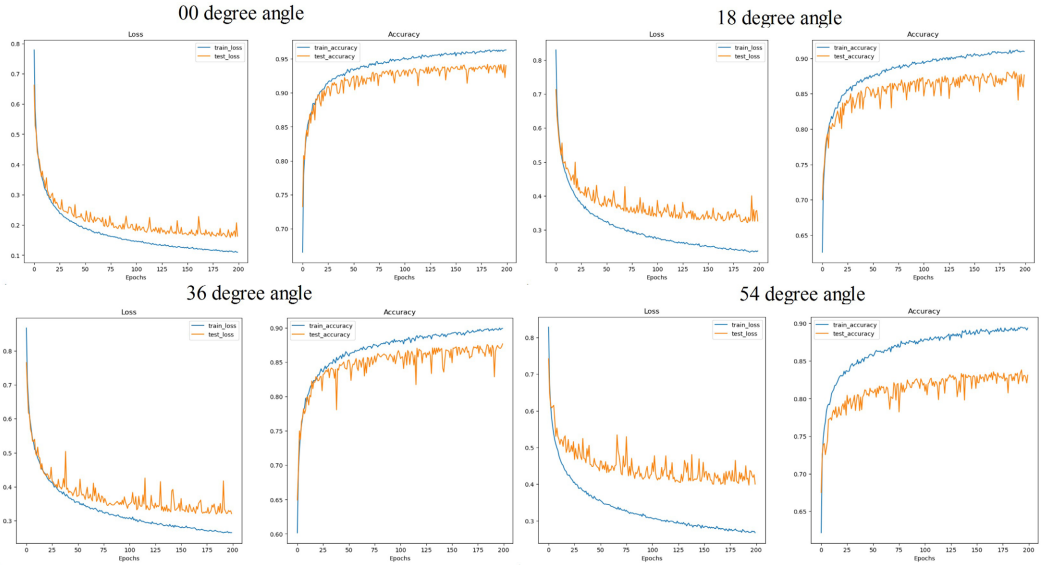

Figure 8. Training curves for the ViT-L/32 model using CASIA-B at a learning rate of 0.001.

However, at higher viewing angles, performance decreased to 83.52% at 36˚ and 80.23% at 54˚. The complete results are provided in **Table 5**, while the training curves are depicted in **Figure 9**.

Table 5. Results for the ViT-L/32 model using CASIA-B at a learning rate of 0.0001.

Angle (˚)	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
0	92.25	90.75	23.54	26.20
18	92.28	90.69	24.51	25.20
36	85.02	83.52	39.78	41.72
54	84.65	80.23	39.24	47.34

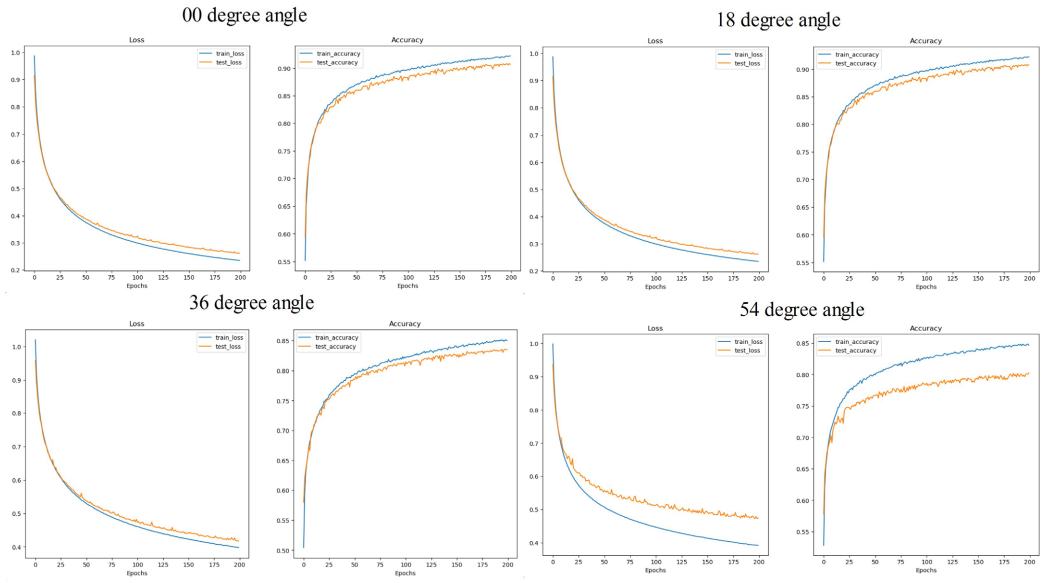

Figure 9. Training curves for the ViT-L/32 model using CASIA-B at a learning rate of 0.0001.

3.8. Performance of ViT-L/32 for Frontal View Analysis Using 00˚ Angle of CASIA-B

The frontal-view analysis using the 000˚ angle of the CASIA-B dataset followed the same experimental protocol described previously for ViT-L/32. Under the gallery vs. NM evaluation setting, ViT-L/32 achieved a training accuracy of 100%, a test accuracy of 96.27%, a training loss of 0.00, and a test loss of 2.15. When evaluated under the gallery vs. CL setting, the model obtained training and test accuracies of 100% and 42.09%, respectively, with corresponding training and test losses of 0.00 and 6.23. For the gallery vs. BG setting, the training accuracy, test accuracy, training loss, and test loss were 100%, 89.98%, 0.00, and 8.21, respectively.

Based on these results, it is evident that ViT-L/32 performs strongly on the NM and BG sets, whereas its performance declines considerably on the CL set. The detailed results are presented in Table 6.

Table 6. Results of ViT-L/32 for frontal view analysis using 00˚ angle of CASIA-B.

Variation	Trn-ACR (%)	Tst-ACR (%)	Trn-LS (%)	Tst-LS (%)
NM	100	96.27	0.00	2.15
CL	100	42.09	0.00	6.23
BG	100	89.98	0.00	8.21
Average	100	76.11	0.00	5.53

The training curves for the frontal view experiments using ViT-L/32 are depicted in Figure 10.

3.9. Performance of ViT-L/32 for Cross View Analysis

For the cross-view analysis of ViT-L/32, 70% of the gait sequences captured at the 000˚ view of the CASIA-B dataset were used for training, while 30% of the sequences captured at the 180˚ view were used for testing. The evaluation considered four metrics: training accuracy, test accuracy, training loss, and test loss. The model achieved a training accuracy of 100%, a test accuracy of 38.84%, a training loss of 0.00, and a test loss of 12.70. The training curves corresponding to this cross-view evaluation using ViT-L/32 are presented in Figure 11.

3.10. Comparative Analysis with Recent HGR Methods

Table 7 is a comparative analysis of the suggested transformer-based framework and the latest HGR methods. The GaitSTAR approach recorded a 76.50% average recognition accuracy at selected angles, and spatial-temporal feature optimisation designs reached around 81.68%. View alignment methods based on GEI achieved average accuracy of less than 80%. The proposed ViT-L/32 model with an average testing accuracy of 87.87 and ViT-B/16 with 86.99 model showed better performance in all tested angles as well.

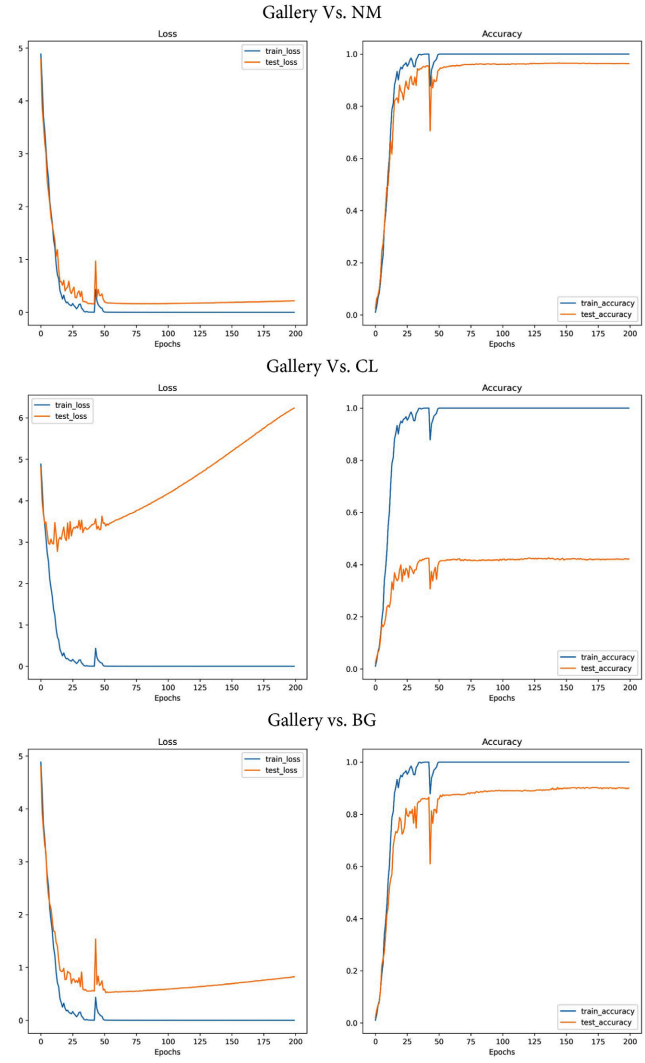

Figure 10. Training curves of ViT-L/32 for frontal view analysis using 00˚ angle of CASIA-B.

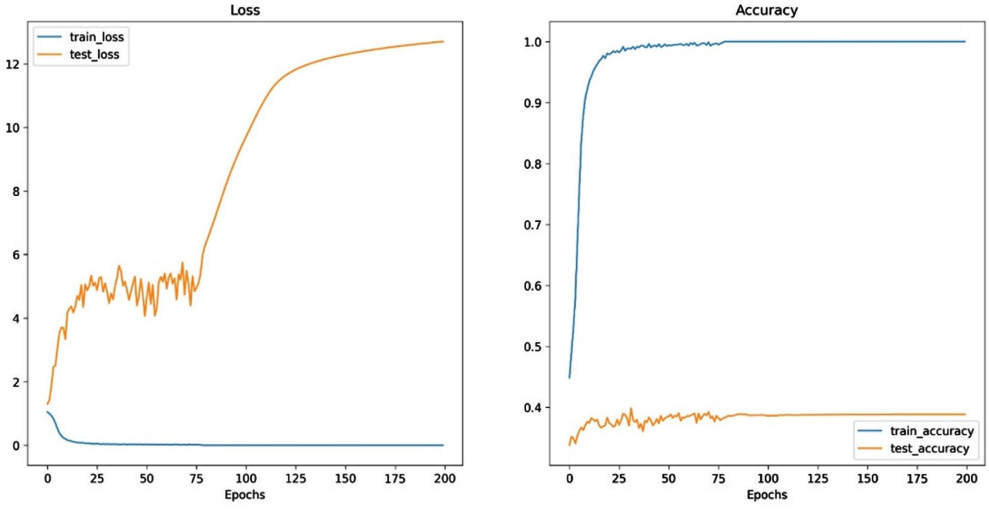

Figure 11. Training curves of ViT-L/32 for cross view analysis.

Table 7. Comparison of the proposed method with recent HGR techniques.

Reference	Year	0˚ (%)	18˚ (%)	36˚ (%)	54˚ (%)	Average (%)
[54]	2024	66.60	75.23	82.60	82.57	76.50
[56]	2022	79.70	81.26	82.76	83.03	81.68
[55]	2020	–	–	–	71.57	71.57
[57]	2019	81.77	78.06	78.60	80.16	79.65
ViT-L/32	—	94.02	87.64	86.69	83.12	87.87
ViT-B/16	—	91.47	86.80	86.28	83.43	86.99

The computational complexity of both models, ViT-B/16 and ViT-L/32, was also evaluated, as summarized in Table 8. The two architectures exhibit distinct parameter counts and FLOP requirements, with ViT-B/16 requiring approximately 44 GFLOPs and ViT-L/32 requiring approximately 55 - 60 GFLOPs. This substantial disparity in computational cost arises primarily from differences in patch size, the number of transformer layers, and the dimensionality of the hidden representations. In comparison, ViT-B/16 demonstrated greater efficiency, owing to its lower FLOP count and reduced inference latency compared to ViT-L/32.

Table 8. Computational complexity of ViT-B/16 and ViT-L/32.

Model	Patch Size	Params (M)	FLOPs (G) (224 × 224 input)
ViT-B/16	16 × 16	~85.8	~44
ViT-L/32	32 × 32	~307	~55 - 60

4. Discussion

The experimental findings indicate that the proposed gait recognition Vision Transformer (ViT)-based framework achieved a high recognition accuracy under various viewing angles of the CASIA-B data. Transformer variants (ViT-B/16 and ViT-L/32) demonstrated both strong performance in perspective changes, with ViT-L/32 models generally outperforming ViT-B/16 in most parameterizations, especially when the learning rate is 0.001.

4.1. Performance Trends across Models and Learning Rates

Models trained at a learning rate of 0.001 performed better than models trained at 0.0001 across all viewing angles, suggesting moderately higher learning rate enabled improved convergence and adaptation to features during fine-tuning. As illustrated in Table 1 and Figure 1, the ViT-B/16 model recorded its best test accuracy (Tst-ACR) at 0˚ of 91.47, where the poor performance reached 83.43 at 54˚. This apparent negative relationship with the horizontal angle reflects the expected difficulty in recognizing cross-view gait since the silhouettes vary dramatically in appearance and motion projection angles.

As presented in Table 2 and illustrated again in Figure 2, the ViT-B/16 model

trained with a 0.0001 learning rate showed a noticeable performance drop. This decline was most evident at the 36˚ and 54˚ viewpoints, where the training accuracy fell below 80%. This underperformance emphasises the importance of selecting appropriate hyperparameters in the application of transformer-based models, and they are also subject to optimisation schedules and data distribution.

Conversely, the ViT-L/32 model exhibited better overall feature discrimination with the highest Tst-ACR of 94.02% at 0˚ with 0.001-learning rate (**Table 3**, **Figure 3**). ViT-L/32 achieved accuracies of over 86% at even more challenging angles (18˚ and 36˚), much higher than the calculated analogs of ViT-B/16. This trend demonstrates the benefit of increased patch sizes and increased transformer layers to capture global gait dynamics and supports related literature indicating the usefulness of transformer attention processes in learning long-range spatial dependencies in vision tasks in intuition [18] [19]. At the reduced learning rate of 0.0001 (**Table 4**, **Figure 4**), ViT-L/32 continued to achieve fairly high performance with Tst-ACRs exceeding 90%.at 0˚ and 18˚.

However, recognition at higher angles (36˚ and 54˚) again declined, illustrating that viewpoint variation remains a fundamental challenge for gait recognition systems even when using advanced architectures. The models were further assessed under frontal-view appearance conditions, and the findings demonstrated that variations in clothing exerted a substantial influence on performance. Moreover, when evaluating view variation, a pronounced degradation in accuracy was observed, indicating a significant sensitivity of the models to changes in viewpoint.

This trend echoes the broader gait recognition literature, where cross-view variation is acknowledged as one of the most persistent sources of error in both appearance-based and model-based frameworks [20].

4.2. Comparative Analysis with State-of-the-Art

When compared with recent HGR methods (**Table 7**), the proposed transformer models achieve superior recognition accuracy across all evaluated angles. The ViT-L/32 model yielded an average accuracy of 87.87%, surpassing techniques such as spatiotemporal attention networks (e.g., GaitSTAR) and hybrid feature optimisation approaches, which have reported average accuracies in the mid-70s to low-80s for the same evaluation protocol. For example, the GaitSTAR method obtained approximately 76.5% average accuracy across the first four angles of CASIA-B, while conventional optimisation and deep learning frameworks ranged between ~71.6% and ~81.7% in recent literature [21]. These results demonstrate that transformer-based global attention aids in extracting discriminative gait features that are less sensitive to covariate variations compared to traditional CNN-based methods and handcrafted feature approaches.

In particular, methods such as Gait-ViT, which apply Vision Transformers to averaged silhouette templates (GEIs), have reported near-perfect CASIA-B performance in controlled settings (e.g., >99% accuracy) [19] [21]. However, such results often stem from tailored preprocessing (e.g., GEI extraction) and evalua-

tion conditions that may not reflect raw sequence inputs. The present study's direct application of ViT models to resized frames (256 × 256) without silhouette averaging demonstrates competitive performance under raw input conditions, highlighting the flexibility and generalisability of transformer architectures for gait representation learning.

4.3. Interpretation of Training Behaviour

The training curves in all the experimental settings are visualised in **Figures 1-4**, which indicate that there are similar patterns of convergence during training and validation stages. It is also worth noting that models educated with the learning rate of 0.001 exhibited a smoother and quicker convergence, and the difference between training and testing losses were less, which suggests improved generalisation. Conversely, the 0.0001 learning rate curves exhibited larger loss gaps and slower rates of convergence, especially ViT-B/16, which reflect the underfitting behavior of more complex visual patterns. Such findings are consistent with the general knowledge on transformer optimisation in vision tasks, where the correct learning rate schedules and patch representations have a strong impact on representation consistency and prediction. Recent studies on gait transformer also propose that patch designs and positional encoding algorithms may have significant impact on recognition performance especially where conditions of cross-view and temporal variation are involved variability [13] [22].

4.4. Limitations and Future Directions

Although the performance is high, the findings validate that viewpoint variation is a core concern of gait recognition, despite the advanced attention-based models. Although ViT-L/32 alleviates these effects to some extent, recognition accuracy declined by approximately 10% across frontal and oblique views, which suggests that alternative strategies, including multi-view fusion, temporal augmentation, or hierarchical attention, can further strengthen them. Future directions could involve studies on self-supervised pretraining with large uncontrolled gait datasets (e.g., GREW) to improve functional generalisation and test performance in unconstrained real-world scenarios.

5. Conclusions

This study introduced a human gait recognition model which was built on ViT-B/16 and ViT-B/32 transfer learning architectures to overcome the problem associated with viewpoint variations, clothing and carrying changes. The two models were tested at various viewing angles of CASIA-B dataset and exhibited optimal recognition performance. ViT-B/16 and ViT-B/32 models recorded 86.99% and 87.87% average testing accuracy, respectively, which surpasses recent gait recognition models.

This high performance is explained by the self-attention mechanism of Vision Transformers, which effectively encodes global and discriminative gait character-

istics in patches of images. The proposed method proved resilient to within-class discrepancies and minimized the computation cost by using pre-trained models. The results highlight the possibilities of transformer-based architectures to efficiently and accurately recognize gaits in practical biometric and surveillance scenarios.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bilal, M., Jianbiao, H., Mushtaq, H., Asim, M., Ali, G. and ElAffendi, M. (2024) Gait-STAR: Spatial-Temporal Attention-Based Feature-Reweighting Architecture for Human Gait Recognition. *Mathematics*, **12**, Article 2458. https://doi.org/10.3390/math12162458

[2] Yao, L., Kusakunniran, W., Wu, Q., Zhang, J., Tang, Z. and Yang, W. (2021) Robust Gait Recognition Using Hybrid Descriptors Based on Skeleton Gait Energy Image. *Pattern Recognition Letters*, **150**, 289-296. https://doi.org/10.1016/j.patrec.2019.05.012

[3] Gao, S., Yun, J., Zhao, Y. and Liu, L. (2022) Gait-d: Skeleton-Based Gait Feature Decomposition for Gait Recognition. *IET Computer Vision*, **16**, 111-125. https://doi.org/10.1049/cvi2.12070

[4] Tyagi, A., Gupta, N., Dwivedi, A., Singh, A. and Srivastava, S. (2026) Human Gait Recognition: A Comprehensive Study Using Deep Learning and Gait Energy Images. In: *Algorithms for Intelligent Systems*, Springer, 427-440. https://doi.org/10.1007/978-981-95-0493-0_33

[5] Sharma, H. and Grover, J. (2018) Human Identification Based on Gait Recognition for Multiple View Angles. *International Journal of Intelligent Robotics and Applications*, **2**, 372-380. https://doi.org/10.1007/s41315-018-0061-y

[6] Shi, L.F., Liu, Z.Y., Zhou, K.J., Shi, Y. and Jing, Y. (2023) Novel Deep Learning Network for Gait Recognition Using Multimodal Inertial Sensors. *Sensors*, **23**, Article 849.

[7] Liu, H., Zhu, Z., Meng, W. and Du, X. (2025) Evaluating Deep Learning in Gait Recognition. In: *Lecture Notes in Computer Science*, Springer, 33-50. https://doi.org/10.1007/978-981-95-3185-1_3

[8] Hans, S., Ranjan, P. and Ismail, S. (2025) Redefining Vision Tasks: The Power of Transformers in Classification, Detection, and Segmentation. In: *Communications in Computer and Information Science*, Springer, 42-53. https://doi.org/10.1007/978-3-031-91340-2_4

[9] Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P. and Wang, L. (2025) Vision Transformers for Image Classification: A Comparative Survey. *Technologies*, **13**, 32. https://doi.org/10.3390/technologies13010032

[10] Li, C., Min, X., Sun, S., Lin, W. and Tang, Z. (2017) DeepGait: A Learning Deep Convolutional Representation for View-Invariant Gait Recognition Using Joint Bayesian. *Applied Sciences*, **7**, 210. https://doi.org/10.3390/app7030210

[11] Aggarwal, H. and Vishwakarma, D.K. (2018) Covariate Conscious Approach for Gait Recognition Based Upon Zernike Moment Invariants. *IEEE Transactions on Cognitive and Developmental Systems*, **10**, 397-407.

https://doi.org/10.1109/tcds.2017.2658674

[12] Alotaibi, M. and Mahmood, A. (2017) Improved Gait Recognition Based on Specialized Deep Convolutional Neural Network. *Computer Vision and Image Understanding*, **164**, 103-110. https://doi.org/10.1016/j.cviu.2017.10.004

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[14] Tang, S., Li, C., Zhang, P. and Tang, R. (2023) SwinLSTM: Improving Spatiotemporal Prediction Accuracy Using Swin Transformer and LSTM. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 13424-13433. https://doi.org/10.1109/iccv51070.2023.01239

[15] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M. and Van Gool, L. (2020) SCAN: Learning to Classify Images without Labels. In: *Lecture Notes in Computer Science*, Springer, 268-285. https://doi.org/10.1007/978-3-030-58607-2_16

[16] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., *et al.* (2023) ResMLP: Feedforward Networks for Image Classification with Data-Efficient Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 5314-5321. https://doi.org/10.1109/tpami.2022.3206148

[17] Xie, S., Girshick, R., Dollar, P., Tu, Z. and He, K. (2016) Aggregated Residual Transformations for Deep Neural Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5987-5995. https://doi.org/10.1109/cvpr.2017.634

[18] Asif, M., Tiwana, M.I., Khan, U.S., Ahmad, M.W., Qureshi, W.S. and Iqbal, J. (2022) Human Gait Recognition Subject to Different Covariate Factors in a Multi-View Environment. Results in Engineering, 15, Article 100556. https://doi.org/10.1016/j.rineng.2022.100556

[19] Mogan, J.N., Lee, C.P., Lim, K.M. and Muthu, K.S. (2022) Gait-ViT: Gait Recognition with Vision Transformer. *Sensors*, **22**, Article 7362. https://doi.org/10.3390/s22197362

[20] Khaliluzzaman, M., Uddin, A., Deb, K. and Hasan, M.J. (2023) Person Recognition Based on Deep Gait: A Survey. *Sensors*, **23**, Article 4875. https://doi.org/10.3390/s23104875

[21] Aman, N., Islam, M.R., Ahamed, M.F. and Ahsan, M. (2024) Performance Evaluation of Various Deep Learning Models in Gait Recognition Using the CASIA-B Dataset. *Technologies*, **12**, Article 264. https://doi.org/10.3390/technologies12120264

[22] Cosma, A., Catruna, A. and Radoi, E. (2023) Exploring Self-Supervised Vision Transformers for Gait Recognition in the Wild. *Sensors*, **23**, Article 2680. https://doi.org/10.3390/s23052680